Probability and cancer clusters

Rachael Hamilton-Keene
Loddon Mallee Integrated Cancer Service
<rhamiltonkeene@bendigohealth.org.au>

Christoper T. Lenard
La Trobe University
<c.lenard@latrobe.edu.au>

Terry M. Mills
Loddon Mallee Integrated Cancer Service
<tmills@bendigohealth.org.au>

Recently there have been several news items about possible cancer clusters in the Australian media. The term "cancer cluster" is used when an unusually large number of people in one geographic area, often a workplace, are diagnosed with cancer in a short space of time. For example, a large number of staff working in one school might be diagnosed with cancer in a given year. In this paper we explore this important health issue using probability theory and in particular the binomial distribution.

Binomial model

The binomial distribution is a standard part of the curriculum in senior mathematics in high school and in first year courses on statistics at university; see, for example, Moore (2007, chap. 13) and Nolan et al. (2006, chap. 11). The key elements of the binomial model are *trial*, *outcome*, and *experiment*. We review these elements in terms of coin tossing.

- A trial consists of tossing a coin.
- The outcome of a trial can be one of two possibilities: heads or tails.
- Let p denote the probability that the outcome of a trial is heads. We assume that this probability is constant and does not vary from trial to trial. We do not assume that the coin is fair; that is, p may have any value between 0 and 1. This allows us to apply the model more broadly.
- The outcome of one trial is independent of the outcome of any other trial. In other words, the result of one toss does not influence the result of another toss.
- An experiment consists of n trials.
- Let X denote the number of trials in the experiment that results in heads. So X could take any of the values 0, 1, 2, … n.

Then X is a random variable, and the probability distribution of X is given by:

$$P(X = t) = \binom{n}{t} p^t (1-p)^{n-t}, \ (t = 0, \ 1, \ \ldots \ n) \qquad (1)$$

We say that X has a binomial distribution with parameters n and p.

Application to the incidence of cancer

Let us apply this model to the diagnosis of cancer. We will focus on Australians in a given age group in a given year.

- A trial consists of observing a person in the given age group for the given year.
- There are two possible outcomes of the trial: either this person is diagnosed with cancer during the year, or not.
- For any trial, assume that the probability that a person is diagnosed with cancer during the year is p, and this probability does not vary from person to person.
- Since cancer is not an infectious disease, we assume that the outcome of one trial is independent of the outcome of any other trial.
- The experiment consists of observing n persons in the given age group over the given year.
- Let X denote the number of persons observed who are diagnosed with cancer during the year.

Since the binomial model seems to fit the situation, X would have a binomial distribution with parameters n and p. Hence, we could calculate the probability function for X using the above formula (1).

Suppose that we are interested in Australians who were between the ages of 25 and 54 in the year 2004. (The ages of members of staff in a school are likely to fall in this range.) There were 8 671 721 persons in Australia between the ages of 24 and 54 in 2004. Of these, 19 572 were diagnosed with some form of cancer in that year. These figures were obtained from the Australian Institute of Health and Welfare (2007). Thus, the probability that a person, who was between 24 and 54 in Australia in 2004, was diagnosed with cancer in that year was $19\,572 \div 8\,671\,721 = 0.2257 \times 10^{-2} \approx 0.0023$.

Consider a particular school that has 60 members of staff between the ages of 25 and 54. Suppose that three or more members of staff were diagnosed with cancer during 2004. Using the binomial formula (1), with $n_1 = 60$ and $p_1 = 0.2257 \times 10^{-2}$ we can show that in 2004, the probability that three or more of these staff would be diagnosed with cancer in 2004 is $0.3573 \times 10^{-4} \approx 0.0004$. This is a very small probability and indicates a very rare event. We might wonder about environmental issues pertaining to that school.

On the other hand, suppose that in Australia in 2004, there were $n_2 = 700$ schools each with 60 members of staff between the ages of 25 and 54. As we saw above, the probability that in any given school, three or more of these staff were diagnosed with cancer is $p_2 = 0.3573 \times 10^{-4}$. Therefore, the probability that in no school were three or more staff diagnosed with cancer in 2004 is $(1 - p_2)700 = 0.7787$. Thus, the probability that, in at least one of the 700 schools, three or more staff were diagnosed with cancer is $1 - 0.7787 = 0.2213$. This is not such a small probability.

So, in a particular school with 60 members of staff, it would be surprising

if three or more staff were diagnosed with cancer in 2004. However, across all 700 schools with 60 members of staff, it is not surprising that one school should have this experience. Realistic data for such calculations can be based on statistics from Australian schools from ABS (2008).

These calculations can be performed using the inbuilt functions of spreadsheets (e.g., BINOMDIST in Excel) and graphing calculators. For example, the TI-83/84 has built-in the cumulative probability distribution function, "binomcdf." Values of n, p and k (separated by commas), can be supplied to obtain the probability $P(X \leq k)$. The probability $P(X = k)$ is found using "binompdf."

Conclusion

We are not suggesting that cancer clusters ought to be ignored by authorities on the basis of an argument from probability. Indeed, they should be investigated thoroughly because the issues associated with such clusters are serious and "complicated" (Schinazi, 2000). The recent report by the Australian Institute of Health and Welfare (2009) is an example of an investigation into mortality and incidence of cancer for certain aircraft maintenance workers. However, the simple model in this paper illustrates a number of points.

In a news item about a cancer cluster at a hospital in New South Wales, a member of the expert panel investigating the situation, Professor Bruce Armstrong, pointed out that this cluster was either caused by some environmental factor, or it "simply occurred by chance" (Kerrin, 2007). The above calculations use the binomial distribution to explain why such a cluster could occur by chance with a reasonable probability.

Similar arguments would apply when considering traffic accidents. If there were a spate of serious accidents at a particular place over a short time period, there would be widespread concern that drastic action should be done about this potential "black spot." Just as a cancer cluster may be due to chance, so too the occurrence of an unusually large number of accidents at one place over a short time may be due to chance.

A goal of the national mathematics curriculum is to enable Australians "to interpret quantitative aspects" of important social issues (National Curriculum Board, 2008). This paper illustrates how students can use the binomial distribution to gain a more thoughtful approach to cancer clusters. It provides a link between senior mathematics and epidemiology, and points the reader to a source of cancer data from AIHW which could be used in the classroom for a variety of exercises or modelling.

Probability theory is a subtle branch of mathematics. To make the above example more concrete, consider this example: if one tosses a fair coin 10 times ($p = 0.5$, $n = 10$), then the probability of obtaining 10 heads is very small (1/1024). However, if 1000 people tossed a coin 10 times, then the probability that someone obtains 10 heads is about 0.62. Although the person who obtained 10 heads would be amazed, it is not amazing that someone obtained 10 heads. This calculation requires careful reflection.

Although coin tossing may appear to be a boring context for teaching and learning, it has the advantage of containing the key aspects of the binomial model without any complicating distractions. Simple models aid our understanding of the essential features of the model. The example shows the importance of coin tossing as a model.

Finally, the paper demonstrates how mathematics should be applied in practice. One starts with a clearly defined model and a good understanding of the assumptions in the model. When putting the model into practice, one has to match the parameters of the model with data in the application. One must check whether the assumptions of the model are reasonable in the circumstances. Fitting a model to a problem in society is one of the exciting parts of applied mathematics.

References

Australian Bureau of Statistics. (2008). *Schools, Australia 2007*. Cat. no. 4221.0. Retrieved 31 October 2008 from http://www.abs.gov.au.

Australian Institute of Health and Welfare [AIHW]. (2009). *Third study of mortality and cancer incidence in aircraft maintenance personnel: Monitoring study of F-111 Deseal/Reseal personnel*. Cancer series no. 45, Cat. No. CAT 41. Canberra: AIHW.

Australian Institute of Health and Welfare. (2007). *Australian cancer incidence and mortality books*. Canberra: AIHW.

Kerrin, L. (2007, 18 July). *Concord hospital investigates cancer cluster*. ABC News 2007. Retrieved 27 October 2008 from http://www.abc.net.au/news/stories/2007/07/18/1982082.htm.

Moore, D. (2007). *The basic practice of statistics (4th ed.)*. New York: W. H. Freeman.

National Curriculum Board. (2008). *National mathematics curriculum: Initial advice*. Retrieved 27 October 2008 from http://www.ncb.org.au/verve/_resources/Mathematics_Initial_Advice_Paper.pdf.

Nolan, J., Phillips, G., Watson, J., Denney, C., Stambulic, S. & Iampolsky, E. (2006). *Maths quest 12: Mathematical methods* (2nd ed.). Brisbane: John Wiley.

Schinazi, R. B. (2000). The probability of a cancer cluster due to chance alone. *Statistics in Medicine, 19*(16), 2195–2198.